DECLARATION

I, Kazunori OKAMURA, of HIRAKI & ASSOCIATES, do solemnly and sincerely declare as follows:

1. That I am well acquainted with the English and Japanese languages and am competent to translate from Japanese into English.

2. That I have executed, with the best of my ability, a true and correct translation into English of Japanese Patent Application No. 2003-348438 filed on October 7, 2003, a copy of which I attach herewith.

This 19th day of December, 2008

K. Okamura

Kazunori OKAMURA

English Translation of Previous Application

Country:	Japan
Type:	Patent
Date of Application:	October 7, 2003
Application Number:	No. 2003-348438
Applicant(s):	Reverse Proteomics Research Institute Co., Ltd.

[Designation of Document] Claims

[Claim 1]

A method for visualizing correlation data concerning two biological events or the correlation data and feature data regarding each event in a matrix format, the method comprising displaying correlation data concerning biological events of the same or different kinds, or the correlation data and feature data regarding each biological event in (a) one of a plurality of prepared data display formats and at (b) one of a plurality of prepared summarization levels, both of which are selected either manually or automatically depending on the number of data items in desired display data.

[Claim 2]

The visualizing method according to claim 1, wherein the plurality of data display formats (a) from which one is selected include: (A) a table data display format having correlation data concerning a pair of events as a single display data unit; (B) a table data display format having correlation data concerning clusters obtained as a result of clustering of events as a single display data unit; and (C) a data display format having the result of statistically processing a set of correlation data as a single display data unit.

[Claim 3]

The visualizing method according to claim 1, comprising a summarization method selected from the plurality of summarization levels (b) that include display or non-display of a data field, reduction of data in a data field of the character type, and reduction of data in a data field of the numeric value type.

[Claim 4]

The visualizing method according to claim 3, wherein the reduction of data in the data field of the character type comprises operations of extracting a part of layers of character information in a layered structure, extracting a keyword from the character data that is registered in advance, and associating the character data with a single sign, letter, or color.

[Claim 5]

The visualizing method according to claim 3, wherein the reduction of data in the data field of the numerical value type comprises operations of rounding a value to an arbitrary significant digit, extracting only the exponential portion of the value, and associating values in a certain range with a color.

[Claim 6]

The visualizing method according to claim 1, wherein the method for automatically selecting the screen display format and the summarization level of data comprises selecting

a pair of a data display format and a data summarization level depending on the number of entries of the correlation data to be displayed on screen and the size of an information display region and an information display unit that are designated in advance, such that a maximum amount of information can be provided.

[Claim 7]

The visualizing method according to claim 1, wherein the correlation data concerning the biological events comprises an interaction between LMW compounds and proteins.

[Claim 8]

A computer-readable recording medium in which a program for causing a computer to implement the visualizing method according to claims 1 to 7 is stored.

[Claim 9]

The computer-readable recording medium according to claim 8, in which the program comprising processes of: constructing correlation data concerning biological events of the same or different kinds, or the correlation data and feature data regarding each biological event in a memory region; and display processing, depending on the number of data items in desired display data, based on (a) one of a plurality of prepared data display formats and (b) one of a plurality of prepared summarization levels, both of which are selected either manually or automatically.

[Designation of Document] Description

[Title of the Invention]

VISUALIZATION METHOD FOR CORRELATED DATA AMONG BIOLOGICAL EVENTS

[Technical Field]

The present invention relates to a method for visualizing correlated data concerning different biological events, particularly information about interactions between substances in living bodies, such as proteins, low-molecular-weight ("LMW") compounds, and DNA, and expression profiles and the like of genes. The invention also relates to a graphical user interface and a visualizing system incorporating the aforementioned method.

[Background Art]

With the completion of the Human Genome Project, information about gene sequences and, moreover, the protein sequences encoded thereby is being comprehensively accumulated. Currently, functional analyses are being actively carried out using such sequence information and proteins, with a view to creating new diagnostic methods and drugs. Knowing protein-protein interactions has a very important meaning in examining the functions of proteins, because the proteins' interactions with other substances in the living body are nothing less than the functions of the proteins. It is believed that, besides protein-protein interactions, correlation information concerning two substances, or, more generally, two events, such as the expression profiles of individual libraries of genes or protein-LMW compound interactions, will shed light on the function of substances in a living body as a system. Therefore, data acquisition has started on a large scale in recent years. It is considered that it becomes increasingly difficult to have an overview of data as a whole and to extract features therefrom, as the volume of data increases. There is the problem that, a large number of detailed references will be required for individual items of data, resulting in the frequent observation of individual sites, as the amount of data increases. Thus, the importance of information visualizing methods for effectively extracting information hidden in large volumes of correlation data is increasing.

In a method for visualizing a large volume of correlation data, a matrix is utilized in which one event is represented in the rows and the other event is represented in the columns, and correlation data concerning the two events is described in the cells where the rows and columns intersect with one another. With regard to expression profiles, a method is generally employed whereby different colors corresponding to expression intensities are displayed in the cells of a matrix. For the visualization of protein-protein interactions, too, a method is employed whereby different colors or shades corresponding to interactions are displayed in the cells of a matrix. For the visualization of protein-LMW compound interactions, too, a method is employed whereby qualitative information, such as "++" or "+," corresponding to interactions is displayed in the cells of a matrix (Patent Document 1).

In the method for displaying the correlation information concerning two events

using a matrix, clustering is generally performed on the basis of patterns in the correlation data of the matrix. By analyzing the nature of the events in the obtained clusters, correspondence between correlation information and the features of each event can be known. Similarly, by sorting the events according to the features of individual events and comparing the resultant correlation information pattern with the features of the events, correspondence between the correlation information and the features of each event can be known. As inferred, in the method of visualizing correlation data using a matrix, it is important that both correlation information patterns and the features of each event can be observed.

Therefore, in an effective method for viewing information, correlation data of a large size in terms of the number of data items thereof is displayed in a matrix, and then characteristic patterns are identified by clustering according to the correlation data patterns or by sorting according to the features of each event. Thereafter, feature quantities regarding the constituent elements of the identified patterns or detailed information about interaction information are accessed, so that it becomes possible to consider the meaning of the obtained patterns. Further, by performing the clustering or sorting in a different manner from the aforementioned clustering or sorting, observing the entirety of the resultant correlation data pattern, and then examining the results to see to what cluster the individual interactions and events that have previously been considered belong, a new discovery could be made. As inferred, by repeating the process of going back and forth between the matrix display of a large quantity of correlation data and the display of individual correlation data, it is believed that new knowledge about correlation data can be discovered.

However, in the conventional method for visualizing correlation data using a matrix, there has been the problem that appropriate information commensurate with the amount of data could not be obtained if the number of data items varies greatly. For example, assume that the number of pixels on a screen is approximately 1000 pixels × 1000 pixels in height and width (30 cm × 30 cm in dimensions). If the data amount is on the order of dozens to 100 items, the number of pixels per cell would be from 10 to dozens of pixels × from 10 to dozens of pixels, which is approximately several mm^2 to 1 cm^2 in dimensions. On this order, the patterns of colors or shades and the individual data points are simultaneously observable.

However, if the data amount increases to the order of several hundreds of items or more, the number of pixels per cell becomes several pixels × several pixels, such that the size of each cell would be 1 mm^2 or smaller. In this case, the cells would be so small that the pattern information becomes complex and, at the same time, it becomes difficult to recognize the individual cells. There would also arise the problem of increased rendering time. Thus, when the data amount reaches the order of several hundreds of items or more as mentioned above, coarse-visualization of patterns can be selected for the description of a single piece of correlation data whereby a certain number of cells or a plurality of cells

corresponding to a cluster are considered together. In this way, the size of each cell can be made to be on the order of several mm to 1 cm × several mm to 1 cm, so that the correlation data pattern and the individual data points can be simultaneously observed. Conventionally, it was previously necessary to conduct this operation manually by the user in a painstaking process.

Conversely, if the size of the rows or columns decreases to dozens or less, the amount of information that can be obtained from the entire screen decreases even though the number of pixels per cell is dozens of pixels × dozens of pixels or more, which corresponds to a fairly large size for each cell; that is, several cm^2. This is due to the fact that the information amount per cell remains at only a level such that it can be represented by colors. If, in order to increase the amount of information obtainable from the entire screen, information about individual cells is to be referred to, it is necessary to access a different information source for each cell. In this case, it has been difficult and troublesome to simultaneously refer to the correlation data pattern and the information about the multiple cells that comprise the pattern.

[Patent Document 1] PCT: WO 02/23199 A2

[Non-Patent Document 1] Advanced Drug Delivery Reviews, 23 (1997) 3-25

[Disclosure of Invention]

[Problem to be solved by the Invention]

As discussed above, in a visualizing method for displaying correlation data concerning two events using a matrix, in order to simultaneously observe a correlation data pattern and information about multiple cells of which the pattern is comprised, it has been necessary to perform some operations, such as coarsely visualizing the correlation data pattern (in which multiple cells are considered together and summarized by clustering and the like), or accessing other sources of information for each cell, depending on the size of the correlation data. In addition, in conventional methods, such operations had to be done manually. As stated with reference to conventional art, in order to discover effective knowledge from a large volume of correlation data, the process of observing the correlation data as a whole and observing a smaller number of items of data in detail has to be repeated. In conventional manual methods, this repetition operation has been done only very inefficiently.

It is an object of the invention to provide means for simultaneously observing a correlation data pattern and information about the multiple cells of which the pattern is comprised in an appropriate manner depending on the variation in the data number size, in a visualizing method for displaying correlation data concerning two events in a matrix.

[Means for Solving Problem]

In order to solve the aforementioned problem, in a screen display system for displaying correlation data concerning two events in a matrix format according to the invention, one of a plurality of data display formats having different levels of integration of data per unit

correlation data that are prepared in advance is automatically selected depending on the variation of the amount of data in terms of the number of items thereof. Also, one of a plurality of display methods having different summarization levels that are prepared in advance for information (about correlation or individual events) regarding individual cells is automatically selected. Then, correlation data and information about individual cells are displayed.

In a typical example of correlation data concerning two events, one event is a protein and the other event is a LMW compound, and the correlation data concerning the events is the intensity of interaction between the protein and the LMW compound. Alternatively, both events may involve proteins and the correlation data concerning the events may be the intensity of protein-protein interaction, or sequence similarity between the proteins. Further alternatively, one event may be gene and the other may be a cDNA library from which the gene derives, and the correlation data concerning the events may be the expression intensity of the gene for each cDNA library. Yet further alternatively, both events may be LMW compounds and the correlation data concerning the events may be structural similarity between the LMW compounds or interaction between them in terms of drug efficacy or side effects.

In accordance with a screen display method of the invention, the data display formats include: (A) a display format (referred to as "individual data display format") in which the elements of correlation data themselves, such as the coupling constants of LMW compounds and proteins, are used as screen display data units; (B) a display format in which groups of a plurality of items of interaction data are used as screen display data units (each group of a plurality of items of interaction data being a cluster obtained by clustering based on correlation data patterns or features of events; hence the format is referred to as a cluster display format); and (C) a display format (statistical display format) in which statistical values of a plurality of items of correlation data are used as screen display data units. The statistical values of correlation data refer to the number of clusters itself or the number of items of related information obtained from a separate data source regarding each element of the cluster, for example.

In accordance with a screen display method of the invention, information about individual cells (information regarding correlation or individual events) is displayed in accordance with a plurality of summarization levels that are set depending on the amount of information. The summarization levels are defined such that they have greater values as the amount of information for expressing a single event decreases.

The plurality of summarization levels defined in accordance with the invention are as follows. When all of the information items that are stored in a data field and that do not overlap one another in meaning are displayed on screen, the data summarization level is defined to be 0 because the data is not summarized. Data formats are defined in advance that correspond to a plurality of summarization levels for different kinds of data fields.

For example, in the display of real number data that includes an exponential portion, the following levels can be adopted:

Summarization level 0: display the field values themselves.

Summarization level 1: display only the exponential portion.

Summarization level 2: classify the values in the exponential portion into five clusters and then display the information using colors associated with the clusters.

Summarization level 3: Display only those that are above a certain threshold value with a color.

In the display of character string data that represents a layered structure, the following levels can be adopted:

Summarization level 0: display the definition of each layer of the layered structure in a staircase-like manner.

Summarization level 1: display the definition of only the upper-most or lower-most layer of the layered structure.

Summarization level 2: display information corresponding to the upper-most or lower-most layer using symbols or colors in a projected manner.

Summarization level 3: display the values of the upper-most layer of the layered structure with associated colors.

A screen display method according to the invention comprises the step of selecting one of the aforementioned multiple data display formats, either automatically or manually, in accordance with the variation of the amount of data in terms of the number of items thereof, the step of selecting one of the aforementioned multiple display methods having different summarization levels for the information (about correlation or individual events) regarding individual cells, either automatically or manually, and the step of displaying the correlation data and information about individual events using the selected data display format and the summarization level.

When the data display format and the summarization level according to the invention are automatically selected, the selection is made such that the amount of information displayed on screen is maintained in the vicinity of a certain value near the maximum amount of information that can be recognized by the user. In other words, the data display format and the summarization level are automatically selected such that all of related information can be displayed within a single screen. However, some scrolling of the screen may be permitted.

In this way, it becomes possible to observe the information about a correlation data pattern and a plurality of cells of which the pattern consists in an appropriate format that is automatically selected depending on the variation in the amount of data in terms of the number of items thereof, without having to manually implement operations such as the coarse visualization of the correlation data pattern or accessing other sources of information

for individual cells depending on the size of correlation data in a visualizing method for displaying correlation data concerning two events in a matrix format. As a result, it becomes possible to implement the operation of repeating the observation of correlation data as a whole and the observation of a relatively small amount of data in detail far more efficiently than possible with the conventional manual operation. Thus, it becomes possible to discover useful knowledge from a large amount of correlation data efficiently.

[Effect of the Invention]

In a method for visualizing the correlation data concerning two biological events in a matrix format, it becomes possible, using the visualizing method of the invention and an interface implementing the visualizing method, to simultaneously observe information about correlation data patterns and the cells of which the patterns are composed in an appropriate display format and at a summarization level that are automatically selected in accordance with the variation in the amount of data, without having to manually implement operations such as causing the correlation data pattern to be coarsely visualized or accessing other sources for information about each cell depending on the size of the correlation data. As a result, regardless of the number of data items to be displayed, it becomes possible to observe the overall picture of the data while the amount of information obtainable from the individual cells is automatically maximized. Thus, it becomes possible to perform the operation of repeating the observation of the correlation data as a whole and the detailed observation of a smaller number of data items far more efficiently than by the conventional manual process. As a result, the process of discovering effective knowledge from a great amount of correlation data can be performed efficiently.

[Best Mode for Carrying Out the Invention]

In the following, embodiments of the invention will be described with reference to the drawings.

[Example 1]

As a correlation between two events, an interaction between substances in a living body, such as proteins, LMW compounds, and DNA is considered. In the following embodiment, data about interactions between "LMW compounds" and "proteins" is handled as the two events that are considered. The term "interaction data" herein refers to information about whether or not there is data about complexes between LMW compounds and proteins in the Protein Data Bank (PDB, http://www.pdb.org), and experimentally measured data showing the degree of binding between LMW compounds and proteins. Feature data about proteins includes the information in various external databases and the calculated results of clustering. It includes, for example, the IDs in SWISSPROT (http://www.expacy.ch/sprot), the clustering results based on amino acid sequence homology, the annotation information based on Gene Ontology (http://www.geneontology.org), and solubilities in a solvent. Feature data about LMW compounds include the names of molecules, molecular weights, therapeutic category, and

other various molecular characteristic values, such as charge distribution, hydrophilic or hydrophobic property, three-dimensional structure, the number of donors or acceptors for hydrogen bond, and the kind and number of functional groups.

With reference to Fig. 1, a flowchart of data visualization is described. A user operation 101 is where data and an action to be performed are selected. Actions include data acquisition 102 and data processing 103. Data acquisition may involve data acquisition by searching a protein-LMW compound interaction database 104 using various search conditions, or data acquisition related to a protein or a LMW compound designated on the display screen from a various-correlation table 105. Data processing may involve clustering entries designated on the display screen, or changing the display scale, for example. The acquired or processed data is handled as display data 106. Then, with regard to the display data, a data display format and a summarization level are determined. The data display format and the summarization level are determined in accordance with a data display format/summarization level determination rule 107 that is prepared in advance, depending on the number of data items in the display data. In accordance with the data display format and the summarization level that have been determined, data screen display 108 is carried out. Conceivable examples of the various correlation table include a protein-protein interaction table, a protein expression profile table, a LMW compound-LMW compound structural similarity table, and a therapeutic or toxicological interaction table.

The main point of the invention, namely, "the data display format and the summarization level are determined in accordance with a data display format/summarization level determination rule that is prepared in advance, depending on the number of data items in the display data," is described in detail below.

First, the data display format is described. Fig. 2 shows an example of a screen display of interaction data concerning LMW compounds and proteins. Labels 201 for LMW compounds are arranged in the vertical direction of the matrix, and labels 202 for the proteins are arranged in the horizontal direction. In a matrix portion 203, there are displayed those experimentally measured coupling constants between proteins and LMW compounds that exceed a certain threshold value, with the bond intensities indicated by different shades. To the left of the compound labels, there are displayed molecular weights 204 as a feature quantity of the compounds. On top of the protein labels, there are displayed the number 205 of alpha helixes and beta strands and clustering information 206 based on protein-protein homology, as feature quantities of the proteins.

With regard to interaction data that is displayed on the screen in a table format, it is also possible to perform clustering based on an interaction data profile or clustering based on the feature quantities of proteins or LMW compounds, and then display the data based on the resultant clustering information.

Clustering using interaction data is conducted by the following method, for

example. A particular LMW compound C_i is considered, and an interaction intensity profile I_{ij} (j=1, ..., N_p, N_p is the number of proteins) of each protein P_j with respect to the LMW compound is considered. Then, the distance D_{ik} between interaction intensity profiles is calculated for all of the LMW compounds on a round robin basis. The distance D_{ik} between interaction intensity profiles of a LMW compound C_i and a LMW compound C_k is calculated in accordance with the following equation, for example:

$$D_{ik} = \sqrt{\sum (I_{ij} - I_{kj})^2}$$

where I_{ij} is the interaction intensity between the LMW compound C_i and the protein P_j.

The sum in the above equation is taken for j=1, ..., N_p.

By providing a threshold value for the round-robin D_{ik} obtained from the above equation, it becomes possible to cluster the LMW compounds. Then, focusing on a single protein P_i, the interaction intensity profile I_{ij} (j=1, ..., N_c, N_c is the number of LMW compounds) of each LMW compound C_j is considered with respect to the protein P_i. By calculating the distance between the interaction intensity profiles of all of the proteins on a round-robin basis, as in the case of the LMW compounds, it becomes possible to cluster the proteins.

Fig. 3 shows the result of actual clustering performed.

LMW compounds are classified into three clusters, and proteins are also classified into three clusters. The results are displayed in an identifiable manner with different shades, with a LMW compound cluster A301, a LMW compound cluster B302, and a LMW compound cluster C303 shown over the labels for the LMW compounds, and with a protein cluster A304, a protein cluster B305, and a protein cluster C306 shown over the labels for the proteins. An average value of coupling constants is internally calculated for each cluster as interaction data, and the clusters are sorted from top to bottom and from left to right in accordance with the averages of coupling constants. Thus, as a general tendency, those cells with high coupling constants (with darker shades) are positioned at the upper-left of the matrix, while those cells with lower coupling constants or whose coupling is less than the threshold value are positioned at the lower-right of the matrix. Such clustering based on the interaction profile allows for the visualization of a cluster 307 consisting of pairs of particular LMW compounds and proteins, and a cluster 308 including many compounds that specifically interact with a single protein, for example. Thus, it becomes possible to adopt an approach, whereby, in an application to the research into drug discovery, a core structure common to the clusters of LMW compounds created on the basis of the interaction profile can be extracted and used as a pharmacophore carrying the functions of a drug as a seed for structural expansion.

Similarly, it is possible to cluster molecular weights into several divisions, or to

classify the numbers of alpha helixes and beta strands of proteins according to certain rules. In this way, it becomes possible to individually rearrange display data for clusters based on molecular weight, clusters based on the number of alpha helixes and beta strands, or clusters based on the homology of amino acid sequences that is calculated in advance. In particular, if a characteristic color pattern of coupling constants appears as a result of rearranging the data according to a certain feature quantity, it can be known that the feature quantity and the coupling constants are closely associated.

Fig. 4 shows the result of rearranging the table in accordance with the result of clustering the LMW compounds based on molecular weight and the proteins based on the amino acid homology. The LMW compounds are classified according to molecular weight into a cluster A401 with a relatively large molecular weight, a cluster B402 with an intermediate molecular weight, and a cluster C403 with a relatively small molecular weight. The data as a whole is sorted in decreasing order of molecular weight. The proteins are classified into a first cluster 404 and a second cluster 405 according to amino acid sequence homology, as displayed on the screen. In the illustrated example, the LMW compounds of cluster B appear to be overlapping a region 406 with a relatively strong interaction in the interaction matrix. On the other hand, there appears no visually recognizable correlation between the clustering result based on the amino acid homology and the interaction intensity. By thus performing clustering with regard to feature quantities and rearranging data in accordance with the result, it could become possible to discover a feature quantity that explains the interaction data well. As a well-known example of feature quantities (molecular characteristic) possessed by a LMW drug, there is the "Rule of five" (the above "Non-Patent Document 1") by Dr. Christopher A. Lipinski. It is believed that, by simultaneously visualizing the clustering result based on feature quantity and the interaction data, it becomes possible to establish rules regarding the feature quantities for explaining certain experimental data, or feature quantities that a LMW compound as a possible target for a particular protein should have.

In the data display in the form of a table as shown in Fig. 3 or 4, each cell in the table corresponds to an interaction between a single protein and a single LMW compound. This is herein referred to as "an individual data display format." The individual data display format, however, has a disadvantage that, as the number of proteins or LMW compounds increases, the table becomes larger and so it becomes more difficult to grasp the data as a whole. Namely, unless the individual cell size is changed in accordance with the increase in the number of data items, the table would not fit within the screen and it would become impossible to view the data as a whole. Conversely, by reducing the individual cell size in the table so as to fit the entire table within the screen, patterns of the interaction data displayed in the cells would become so small that it becomes difficult to recognize their features. In order to allow the interaction patterns in the table as a whole to be recognized at a glance even if the number of data items increases, therefore, it is herein made possible

to display the information using the individual clusters in Fig. 3 or 4 as a single cell on the table. This is herein referred to as "a cluster display format."

Fig. 5 shows an example of the cluster display format. In labels 501, cluster numbers are entered. As feature quantities, the number 502 of elements that belong to a cluster and a list 503 of elements that belong to a cluster are shown. In a matrix portion 504, average values of measurement data for each cluster are displayed with different shades, and the number of elements of which a cluster is comprised is indicated by a numerical value. Information display can be switched between the individual data display format and the cluster display format. Rearrangement of rows or columns or other operations such as deletion in one display format is reflected in the other display format. Because in the cluster display format, clusters are formed by similar proteins or similar LMW compounds, representative data can be visualized without fail. By controlling the number of clusters at the same time, the number of rows or columns of the displayed table can be controlled even when the number of interaction data items is large.

As a supplementary information display format to the individual data display format and the cluster display format, there is "a statistical quantity display format." This is a format wherein statistical calculations are performed on all or part of the data and the resultant average values or standard deviations, for example, are displayed, or wherein the number of data items extracted from a different data source is displayed. In the statistical quantity display format, it is possible to have an overview of the data regardless of the number of the interaction data items. When the number of data items increase, it becomes difficult to recognize the interaction pattern of the table as a whole at a glance. In such cases, the statistical quantity display format proves very effective for grasping the overall picture of the data.

In accordance with the invention, a plurality of display formats are prepared. In addition, different levels of summarization of information to be displayed in the individual cells of a matrix are prepared, and an appropriate one can be selected depending on the number of data items.

For the display of the interaction data between proteins and LMW compounds, four levels of summarization (0 to 4) are prepared. At summarization level 0, all of the information stored in the database and the statistical quantities and the like calculated therefrom are displayed. At summarization level 1, character data of up to 64 letters per cell, signs, or colors can be displayed. It is also possible to display information consisting of 64 letters or fewer in a text field of the database, or even longer information as long as it can be reduced to 64 letters or fewer. At summarization level 2, character data consisting of up to 8 letters per cell, signs, or colors can be displayed. At summarization level 3, no character data is displayed and instead the entire information is represented by colors.

In actual implementation, the information display at summarization level 0 is performed on a free format basis. At summarization level 1, the size of each cell is

defined as consisting of 60 pixels vertically and 120 pixels horizontally, in which a region is secured for displaying 16 letters × 4 rows of text. At summarization level 2, the size of each cell is defined as consisting of 20 pixels vertically and 60 pixels horizontally, in which a region is ensured for displaying 8 letters × one row of text. At summarization level 3, the size of each cell is defined as consisting of 5 pixels vertically and 5 pixels horizontally. In principle, it is possible to reduce the single cell size down to 1 pixel × 1 pixel. However, the cell size is selected such that individual data items can be manipulated using the mouse.

Screen display at these four summarization levels can be switched. Fig. 6 shows examples of the screen display of information at the four summarization levels in the individual data display format.

In a screen display 601 at summarization level 0, interaction data, LMW compound data, and protein data are displayed in detail. The display format is free, so that it is possible to display and manipulate the structure of a protein or LMW compound, for example.

In a screen display 602 at summarization level 1, keys for accessing various external protein-related databases, the names and therapeutic effects of LMW compounds, and detailed values of measurement data about interaction, for example, are displayed.

In a screen display 603 at summarization level 2, the displayable character data is limited to 8 letters, so that limited information, such as the labels for identifying the row or column and major values of measurement data about interaction are displayed.

In a screen display 604 at summarization level 3, the values taken by the individual cells are converted into color information when displayed. In this way, similar data can be visually recognized from the color pattern.

For the data items that are selected, it is necessary to create a rule regarding how the information is to be summarized at a given summarization level. A basic rule is such that, at summarization level 0, all of the information is displayed; at summarization levels 1 and 2, information is displayed depending on the length of the character; and at summarization level 3, information is displayed in terms of colors. In accordance with this basic rule, a detailed summarization rule must be defined for each of the data items that exist in the database.

For example, Fig. 7 shows a summarization rule determination table for a LMW compound feature table. Information is given as to which data item 702 in the fields of the table is to be processed in which location 703 and according to what summarization rule 704 for screen display, depending on summarization level 701.

If the field name does not appear in the summarization rule determination table, this means that that field will not be displayed. If the summarization rule is "as is" 705, the data stored in the database will be displayed as is. In another example, if the summarization rule is "color (200, 300, 400, 500)" 706, different colors will be displayed for the five cases of the values of less than 200, 200 or more and less than 300, 300 or more

and less than 400, 400 or more and less than 500, and 500 or more. Such a summarization rule determination table needs to be provided in each of the tables in the database.

So far, three data display formats and four data summarization levels have been described. By combining these, data can be visualized in a variety of different perspectives. The invention is characterized by the function whereby, as the user selects desired information, an optimum data display format and data summarization level are automatically determined in accordance with the number of data items in the selected information.

Examples of input data necessary for the automatic determination of the data display format and the data summarization level when visualizing interaction data concerning proteins and LMW compound, for example, include the number P of proteins, the number C of LMW compounds, the number Pc of protein clusters, the number Cc of LMW compound clusters, and parameters x (height) and y (width) of the information display region on the screen. When there are multiple kinds of clusters, the number of clusters registered as an initial setting is used.

Fig. 8 shows a table of rules for determining the data display format and the data summarization level. Condition 801 is viewed one by one from the top and a display format 802 and a summarization level 803 described in the line where the condition is satisfied are adopted. If the condition is not satisfied, the condition in the next line is viewed. G, R, Gc, and Rc are the values defined in Fig. 8. The table is described below.

If $P \times C$ (corresponding to the number of cells in the display screen) is smaller than a predetermined value (3 in the example), summarization level 0 is used for the individual data display.

If $P \times C > 3$, $G \leq 11$, and $R \leq 11$, the number P of proteins and the number C of LMW compounds would be both 2 or more and 9 or less when the number of displays of feature quantities in the column direction and the number of displays of feature quantities in the row direction are both 1. In this case, the summarization level 1 is used, so that the size of a single cell would be 60 pixels vertically and 120 pixels horizontally. Thus, in the information display region of 450 pixels vertically and 900 pixels horizontally, the display size for the entire data would be 240 pixels vertically × 480 pixels horizontally to 660 pixels vertically × 1320 pixels horizontally, which is within 1.5 × 1.5 times the entire information display region.

As the number P of proteins and the number C of LMW compounds increase, the summarization level is increased from 2, 3, and so on sequentially in accordance with Fig. 8. If the numbers P and C further increase, the display format is switched to the cluster display format, and the summarization level is increased from 1, 2, 3, and so on, as the number Pc of the protein clusters and the number Cc of LMW compound clusters increase.

The conditions with regard to G, R, Gc, and Rc for the switching of display format and summarization level are set such that the display size for the entire data would be within

1.5 × 1.5 times the entire information display region. If a generalized standard that information regarding the entire data be displayed within n × m times the data display region is to be satisfied, the following generalized condition can be used for the determination of the data display format and summarization level:

$x \times n \leq P$ (or Pc) and $y \times m \leq C$ (or Cc)

In this way, it becomes possible to display the entire data, or an amount of data that is a certain multiple of the entire data, within the information display region. Also, by increasing or decreasing the summarization level depending on the decrease or increase in the number of data items, it becomes possible to display a maxim amount of information within a cell that can be recognized at a glance. Thus, it becomes possible to observe the entire picture of the data while the amount of information that can be obtained from each cell is maximized, regardless of the number of data items to be displayed.

In the process of discovery of targets for the creation of new drugs, it is extremely important to visualize the interaction between proteins and LMW compounds and, at the same time, to acquire information about other relevant biological interactions, put the information in order comprehensibly, and understand them. Examples of the relevant biological interactions include interactions between LMW compounds regarding drug efficacy or toxicity, interactions between proteins, and information regarding proteins and expression. In accordance with the invention, it is possible to acquire those related information and then display them depending on the number of data items acquired and in accordance with the above-described determination rule regarding the display format and summarization level.

Related information is acquired in the following manner. A cell region of interest in a displayed data table is selected, and then the LMW compound IDs and the protein IDs that belong in the cell region are extracted. A related data table is searched for these IDs, and the information associated with the retrieved IDs are extracted from the related data table.

Fig. 9 shows a concrete method for extracting related information. When two items of a protein-LMW compound interaction table 901, namely, (C5, P12) and (C9, P12) are considered, those of the cells of a protein-protein interaction table 902, which standardizes the protein-protein coupling strength against a maximum value of 100, and of a protein-expression table 903, which indicates the quantitative expression amounts of protein in an expression library, are extracted that have P12 as the protein ID and for which data exists. Similarly, those of the cells of a LMW compound-LMW compound interaction table 904, in which data regarding the presence or absence of effects due to multiple drug use between LMW compounds is stored, are extracted that have C5 and C9 as IDs and for which data exists.

The result of extraction of related information is displayed as arranged for each table from which the information has been extracted, as shown in Fig. 10. When the user

selects a table he or she wishes to see, an information display format and summarization level are automatically set depending on the number of hits, and information is displayed in the display format and at the summarization level that have been set. Related information can also be acquired from part of the information thus displayed. Thus, the invention allows the visualization of multi-dimensional interaction data by efficiently tracking the links between the one-to-one interaction data.

In accordance with an interface implementing the visualization method of the invention, part of the information displayed on screen is selected, an action selected from a plurality of actions is performed on the selected data, and the information obtained as a result of the action is displayed on screen. Fig. 11 shows an example of the user interface. In addition to a display mode change button 1101, a summarization level change button 1102, and a related information acquisition button 1103, there are provided a function group 1104 related to actions, such as replacement, rearrangement, clustering, and deletion of rows or columns, and a function group 1105 related to the selection of characteristic rows or columns or those rows or columns as representative subsets. A mouse-operated action is assigned to each of the cells displayed on screen in the table format, allowing a row or column to be selected, or long character string data that cannot be displayed within the cell to be displayed on a related information display screen 1106.

[Brief Description of Drawings]

[Fig. 1] Fig. 1 shows a flowchart of data visualization.

[Fig. 2] Fig. 2 shows an example of the screen display of interaction data regarding LMW compounds and proteins.

[Fig. 3] Fig. 3 shows an example of the screen display of data sorted according to the result of clustering based on an interaction data profile.

[Fig. 4] Fig. 4 shows an example of the screen display of data sorted according to the result of clustering based on feature quantities in the rows and columns.

[Fig. 5] Fig. 5 shows an example of the display of information in a cluster display format.

[Fig. 6] Fig. 6 shows an example of the screen display of information in an individual data display format at four summarization levels.

[Fig. 7] Fig. 7 shows rules for the determination of the data display format and the data summarization level.

[Fig. 8] Fig. 8 shows a summarization rule determination table for a LMW compound physical property table.

[Fig. 9] Fig. 9 shows the outline of a method for extracting related information.

[Fig. 10] Fig. 10 shows the result of extraction of related information.

[Fig. 11] Fig. 11 shows an example of the screen of a user interface implementing the invention.

[Explanations of Letters or Numerals]

101: user operation, 102: data acquisition, 103: data processing, 104: protein-LMW compound interaction database, 105: tables of various correlations, 106: display data, 107: data display format and summarization level determination rules

201: LMW compound labels, 202: protein labels, 203: matrix portion, 204: molecular weight, 205: number of alpha helixes and beta strands, 206: clustering information based on homology

301: LMW compound cluster A, 302: LMW compound cluster B, 303: LMW compound cluster C, 304: protein cluster A, 305: protein cluster B, 306: protein cluster C, 307: cluster consisting of pairs of particular LMW compounds and proteins, 308: cluster consisting of pairs of compounds and a single protein that have a specific interaction

401: cluster A with a relatively large molecular weight, 402: cluster B with an intermediate molecular weight, 403: cluster C with a relatively small molecular weight, 404: cluster 1 based on the homology of amino acid sequences, 405: cluster 2 based on the homology of amino acid sequences, 406: region with a relatively strong interaction

501: labels, 502: number of elements that belong to a cluster, 503: list of elements that belong to a cluster, 504: matrix portion

601: screen display at summarization level 0, 602: screen display at summarization level 1, 603: screen display at summarization level 2, 604: screen display at summarization level 3

701: summarization level, 702: data items, 703: location, 704: summarization rules, 705: rule "As is", 706: rule "colors (200, 300, 400, 500)"

801: condition, 802: display format, 803: summarization level

901: protein-LMW compound interaction table, 902: protein-protein interaction table

903: protein-expression table, 904: LMW compound-LMW compound interaction table

1101: display mode change button, 1102: summarization level change button, 1103: related information acquisition button, 1104: functions related to actions, 1105: functions related to selection, 1106: related information display area

[Designation of Document] Abstract

[Abstract]

[Problems of the Invention] The amount of work required to repeat the operation of observing the correlation data as a whole and the detailed observation of a small amount of data is reduced in a method for visualizing correlation data concerning two events in a matrix format.

[Means to Solve the Problems] One of a plurality of data display formats with different units of correlation data that are prepared in advance is automatically selected depending on the variation of the amount of data in terms of the number of items thereof. One of a plurality of display methods with different summarization levels that are prepared in advance is also automatically selected for information (about correlation data or individual events) regarding individual cells. Information about the correlation data and individual cells is then displayed.

[Representative Drawing] Fig. 1

FIG. 1

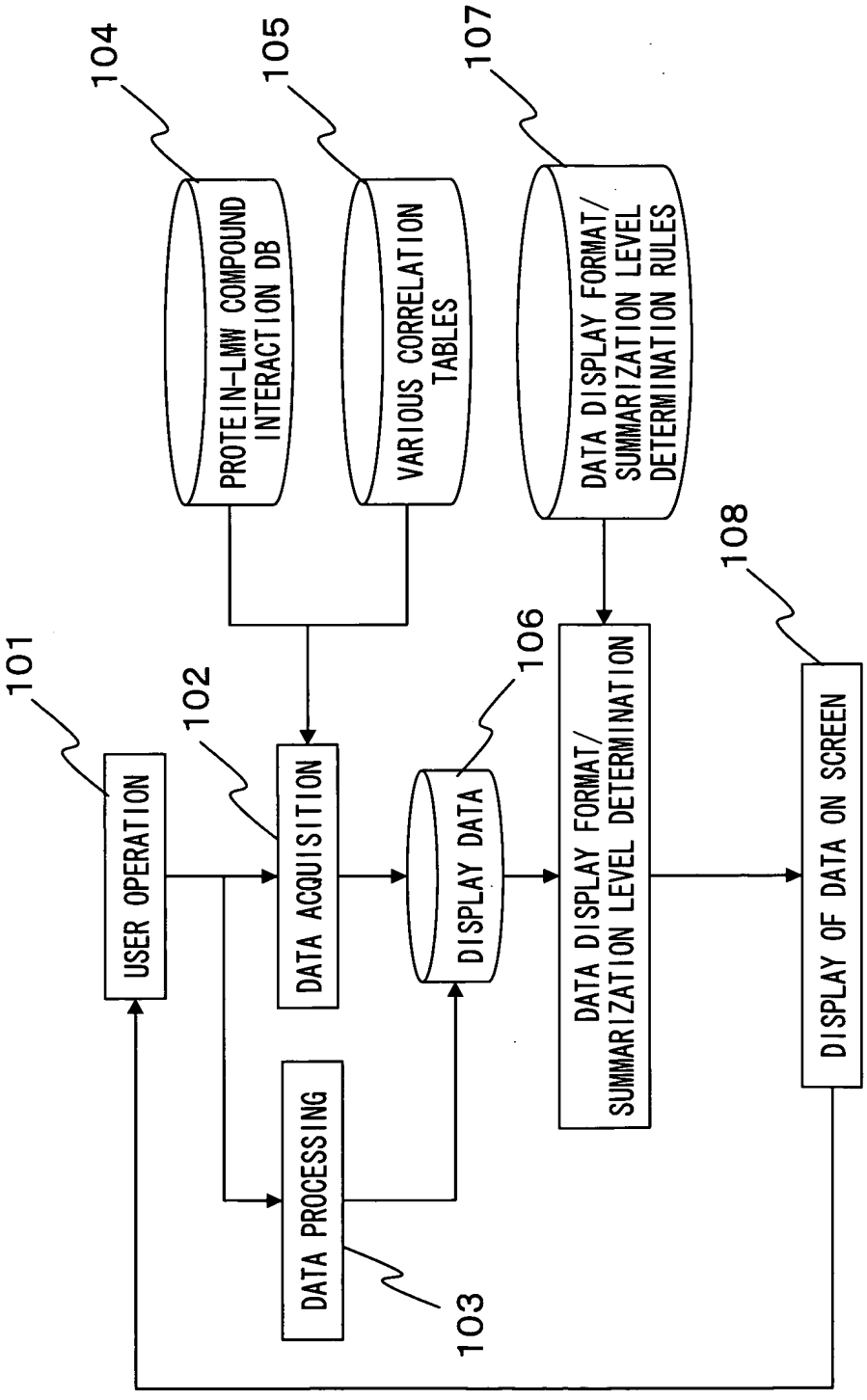
104
105
107
PROTEIN-LMW COMPOUND INTERACTION DB
VARIOUS CORRELATION TABLES
DATA DISPLAY FORMAT/ SUMMARIZATION LEVEL DETERMINATION RULES
101
102
106
108
USER OPERATION
DATA ACQUISITION
DISPLAY DATA
DATA DISPLAY FORMAT/ SUMMARIZATION LEVEL DETERMINATION
DISPLAY OF DATA ON SCREEN
DATA PROCESSING
103

FIG. 2

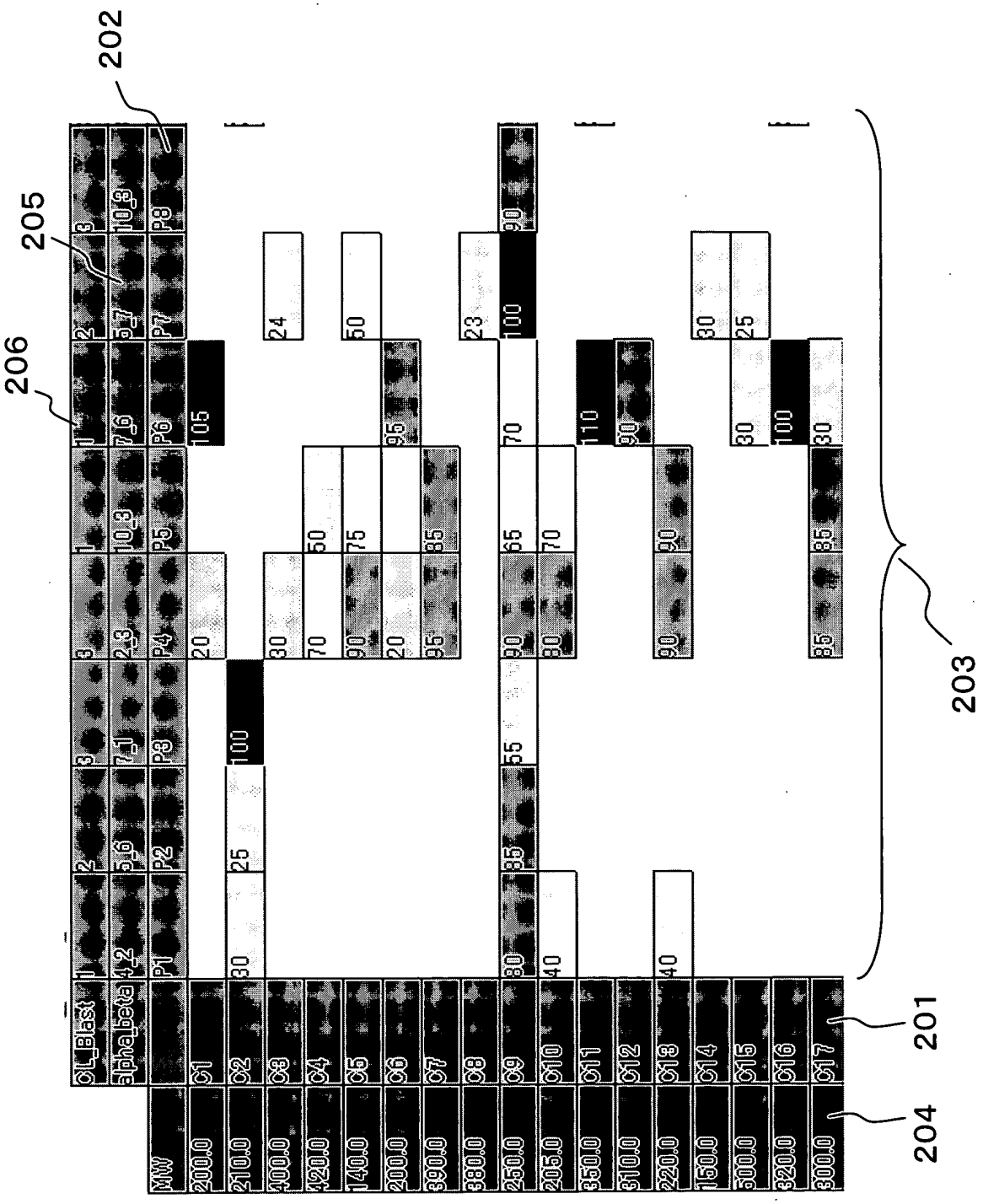
202
205
206
203
201
204
CL_Blast
alpha_beta
MW
P1
P2
P3
P4
P5
P6
P7
P8
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16
C17
200.0
210.0
400.0
420.0
140.0
200.0
390.0
380.0
250.0
205.0
350.0
310.0
220.0
150.0
300.0
320.0
300.0

FIG. 3

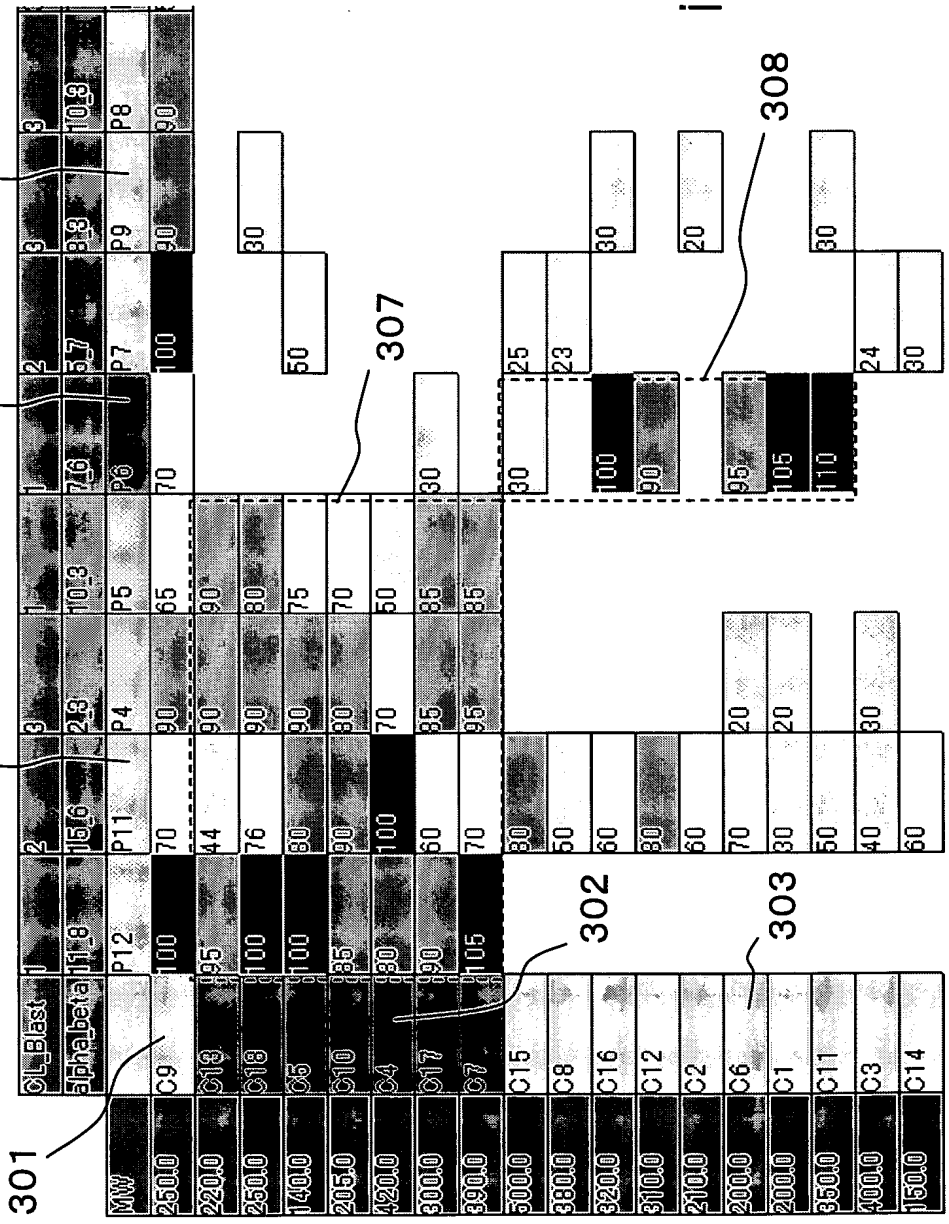
301
302
303
304
305
306
307
308
i
CL_Blast
alpha_beta
MW
250.0
220.0
250.0
140.0
205.0
420.0
300.0
390.0
500.0
380.0
320.0
310.0
210.0
200.0
200.0
350.0
400.0
150.0

FIG. 4

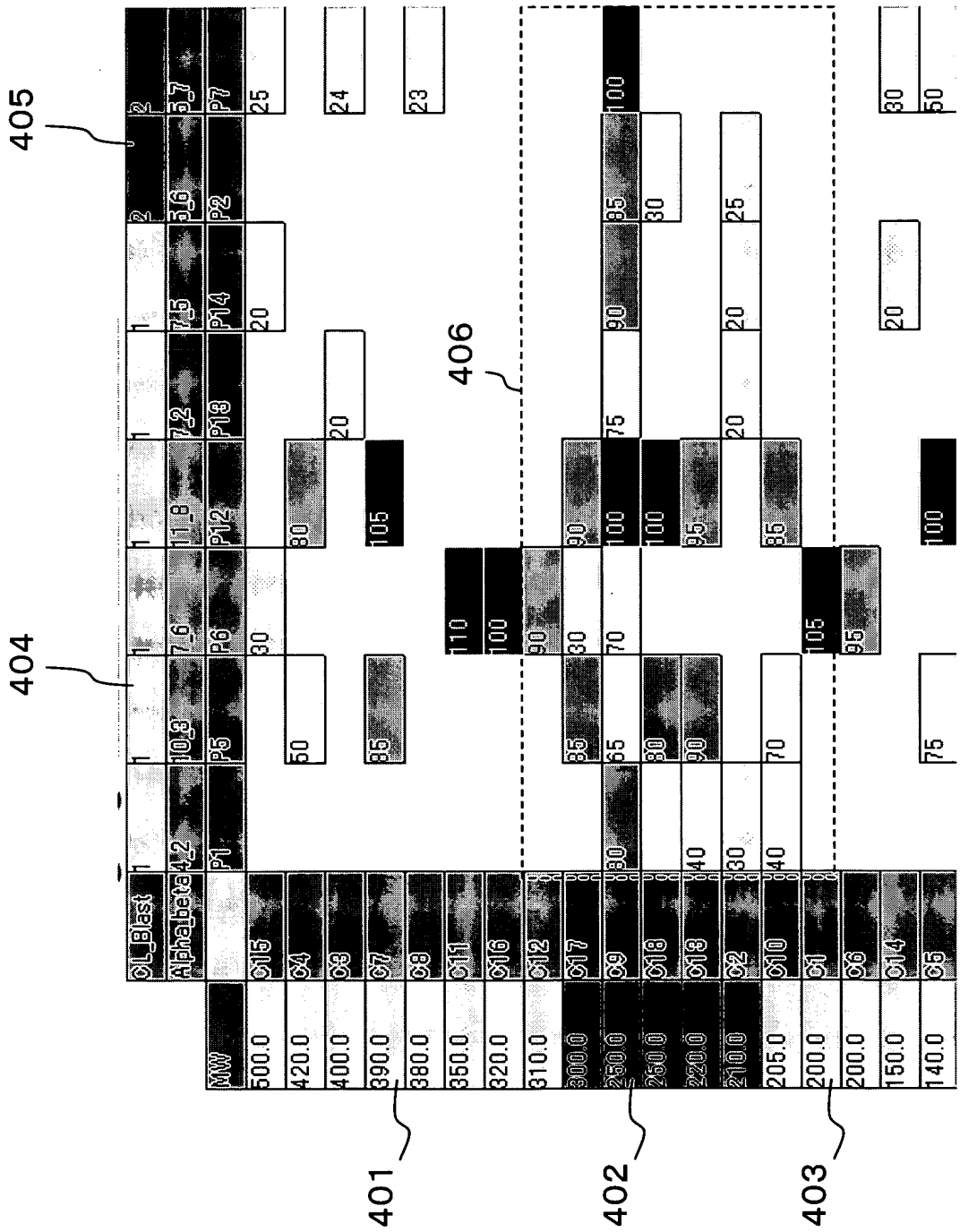
405
404
406
401
402
403
CL_Blast
Alphabeta_2
NW
500.0
420.0
400.0
390.0
380.0
350.0
320.0
310.0
300.0
250.0
260.0
220.0
210.0
205.0
200.0
200.0
150.0
140.0

FIG. 5

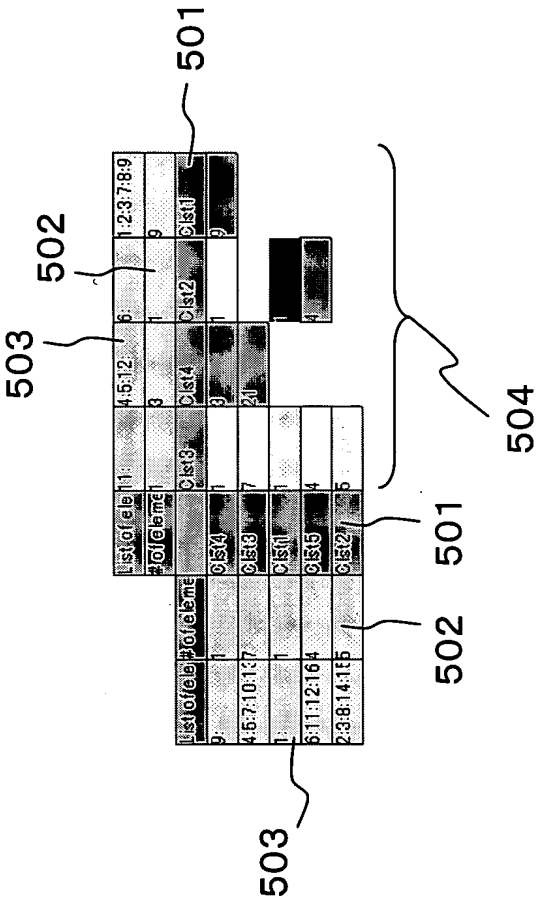

FIG. 6

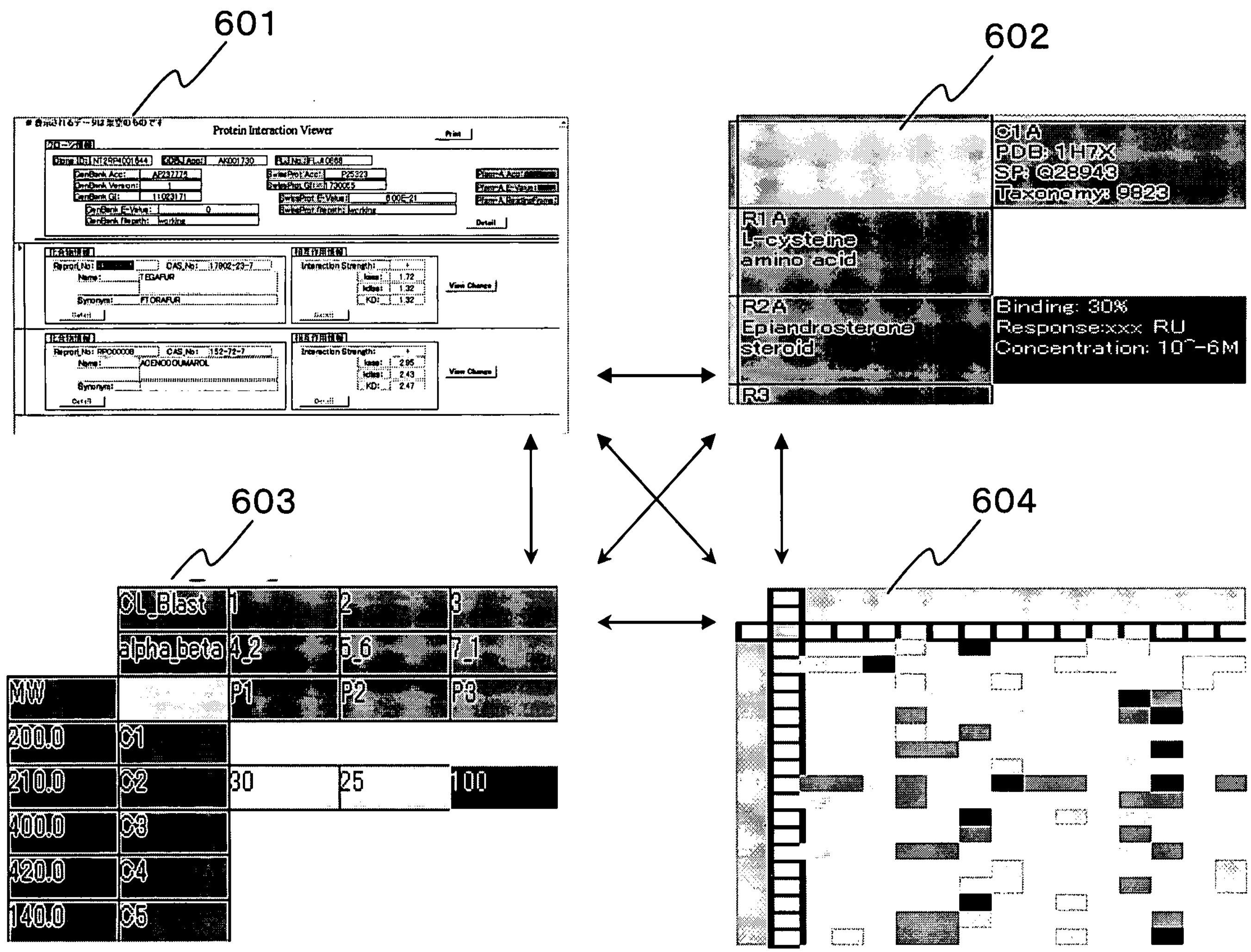
601
Protein Interaction Viewer
Print
Detail
View Change
602
C1A
PDB: 1H7X
SP: Q28943
Taxonomy: 9823
R1A
L-cysteine
amino acid
R2A
Epiandrosterone
steroid
R3
Binding: 30%
Response:xxx RU
Concentration: 10^-6M
603
CL_Blast
alpha_beta
MW
1
2
3
4_2
5_6
7_1
P1
P2
P3
200.0
210.0
400.0
420.0
140.0
C1
C2
C3
C4
C5
30
25
100
604

FIG. 7

SUMMARIZATION LEVEL (701)	DATA ITEM (702)	LOCATION (703)	SUMMARIZATION RULE (704, 705)
1	NAME OF COMPOUND	LABEL	AS IS
1	MOLECULAR WEIGHT	FEATURE QUANTITY DESCRIPTION CELL	AS IS
1	PHYSICAL PROPERTY CLUSTER NUMBER	FEATURE QUANTITY DESCRIPTION CELL	AS IS
1	NAME OF DRUG EFFICACY CLUSTER	FEATURE QUANTITY DESCRIPTION CELL	AS IS
2	COMPOUND ID	LABEL	LAST 5 DIGITS
2	MOLECULAR WEIGHT	FEATURE QUANTITY DESCRIPTION CELL	ROUND TO WHOLE NUMBER
2	PHYSICAL PROPERTY CLUSTER NUMBER	FEATURE QUANTITY DESCRIPTION CELL	LAST 5 DIGITS
2	DRUG EFFICACY CLUSTER NUMBER	FEATURE QUANTITY DESCRIPTION CELL	LAST 5 DIGITS
2	NAME OF DRUG EFFICACY CLUSTER	INFORMATION DISPLAY SEPARATE SCREEN	LINK FROM DRUG EFFICACY CLUSTER NUMBER; AS IS
3	COMPOUND ID	INFORMATION DISPLAY SEPARATE SCREEN	LINK FROM LABEL; AS IS
3	MOLECULAR WEIGHT	FEATURE QUANTITY DESCRIPTION CELL	COLORS (200, 300, 400, 500)
3	PHYSICAL PROPERTY CLUSTER NUMBER	FEATURE QUANTITY DESCRIPTION CELL	COLORS (DIFFERENT COLOR FOR EACH NUMBER)
3	NAME OF DRUG EFFICACY CLUSTER	FEATURE QUANTITY DESCRIPTION CELL	COLORS (DIFFERENT COLOR FOR EACH NUMBER)

706

FIG. 8

801 CONDITION	802 DISPLAY FORMAT	803 SUMMARIZATION LEVEL
P×C≦3	INDIVIDUAL DATA DISPLAY	0
G≦11 & R≦11	INDIVIDUAL DATA DISPLAY	1
G≦34 & R≦22	INDIVIDUAL DATA DISPLAY	2
G≦135 & R≦270	INDIVIDUAL DATA DISPLAY	3
Gc≦11 & Rc≦11	CLUSTER DISPLAY	1
Gc≦34 & Rc≦22	CLUSTER DISPLAY	2
Gc≦135 & Rc≦270	CLUSTER DISPLAY	3
OTHERS	STATISTICAL DISPLAY	0

DEFINITION	
	G: P + NUMBER OF DISPLAYS OF PHYSICAL PROPERTY IN COLUMN DIRECTION + 1
	R: C + NUMBER OF DISPLAYS OF PHYSICAL PROPERTY IN COLUMN DIRECTION + 1
	Gc: Pc + NUMBER OF DISPLAYS OF PHYSICAL PROPERTY IN COLUMN DIRECTION + 1
	Rc: Cc + NUMBER OF DISPLAYS OF PHYSICAL PROPERTY IN COLUMN DIRECTION + 1

FIG. 9

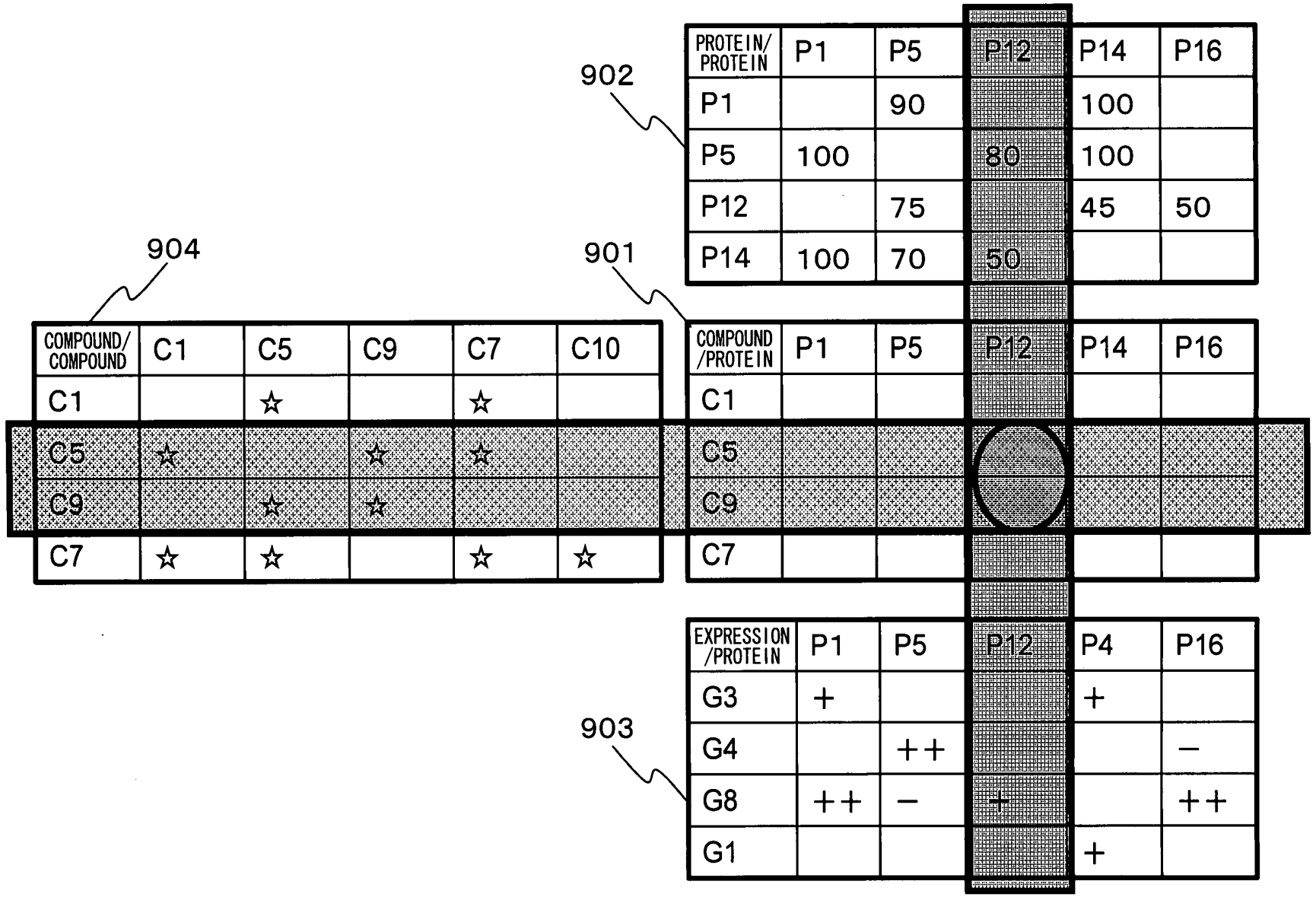
902
PROTEIN/PROTEIN
P1
P5
P12
P14
P16
P1 90 100
P5 100 80 100
P12 75 45 50
P14 100 70 50
904
901
COMPOUND/COMPOUND
C1
C5
C9
C7
C10
COMPOUND/PROTEIN
903
EXPRESSION/PROTEIN
P4
G3 + +
G4 ++ −
G8 ++ − + ++
G1 +

FIG. 10

RELATED INFORMATION ABOUT (R5, C12), (R9, C12)
FROM PROTEIN-EXPRESSION TABLE: xx ITEMS
FROM PROTEIN-PROTEIN INTERACTION TABLE: yy ITEMS
FROM LMW COMPOUND-LMW COMPOUND INTERACTION TABLE:
zz ITEMS

FIG. 11